Compendium of Computer Science
ISSN Print: 3079-5362
ISSN Online: 3079-5370
https://doi.org/10.63313/CS.8002

Social Structure-Based Methodology for User Identity Matching Across Social Networks

Guanlong Zhu
School of Computer Science and Technology, Qingdao University, Qingdao, China
Email: kanmurinana@gmail.com

How to cite this paper: Zhu, G. (2025). Social Structure-Based Methodology for User Identity Matching Across Social Networks. Compendium of Computer Science, 1(1), 23-32. ISSN Print: 3079-5362; ISSN Online: 3079-5370.
https://doi.org/10.63313/CS.8002
Published: 2025-03-25

Abstract

Cross-social-network user matching technology can identify the same individual across different social networks, enabling cross-network data integration. This allows for precise user profiling, which facilitates applications such as user recommendation and influence maximization. In such tasks, the social structure of users plays a critical role. However, due to the vast and complex nature of social network data, issues such as data sparsity and heterogeneity across net-works arise. Additionally, as the depth of network layers increases, extracted features tend to degrade and become homogeneous. To address these challenges, this paper proposes a Social Network Structure-based User Matching (SSUM) method. The approach introduces a subgraph structure extraction module, which, combined with the original network topology, is fed into a Transformer layer to further extract global features. This mitigates over-smoothing and over-squeezing caused by excessive layer depth, thereby obtaining vector rep-resentations of user topological structures. Subsequently, word-level vectors of user information and text-level vectors of user-generated content are extracted as user attribute features. These are concatenated to form the final user embed-ding representation. User matching is then transformed into a ranking problem by calculating Manhattan distances. Furthermore, to enhance model perfor-mance, user relationships in the dataset are pre-completed. Experimental results on real-world social network datasets demonstrate that the proposed method effectively extracts user features, improves matching accuracy, and outperforms comparative models.

Keywords

Social Networks, User Matching, Network Embedding, Transformer

1. Introduction

The continuous advancement of internet technology has laid a solid foundation for the rise of social networks, leading to the emergence of numerous online social platforms such as Xiaohongshu, Douban, and Weibo in China, as well as Facebook and Twitter abroad. As social platforms diversify in functionality, an increasing number of users choose to register accounts on multiple platforms to access different communities, fulfill varied needs, and experience distinct features. These users leave behind rich social data across different platforms. However, since these platforms are often independent of one another, information isolation arises. Collecting and analyzing user information across these platforms can enable data fusion, offering broad application value in areas such as recommendation systems, social network analysis, and cybersecurity.

Nevertheless, cross-social-network user matching faces numerous challenges due to the heterogeneity, sparsity, and dynamic nature of social network data. For instance, different platforms may adopt varying data formats and structures, and user information across platforms may be incomplete or noisy. Additionally, privacy protection issues limit the availability of user data, as many users employ different identities across platforms, further complicating the matching process.

In recent years, researchers have proposed various cross-social-network user matching methods, which can be broadly categorized into unsupervised, semi-supervised, and supervised approaches based on the availability of labeled data. Unsupervised methods match users based on attributes, social relationships, or content features without relying on labeled data, though their matching accuracy is limited. Early methods were predominantly unsupervised. For example, Gao et al. proposed the CNL method [1], which utilizes multiple user attributes and designs an incremental approach for identity matching. Li et al.[2] transformed the unsupervised user identity matching problem into a projection function learning problem, minimizing the distribution distance between users across platforms using the Earth Mover's Distance as a metric. Fact Embedding [3] is a recent unsupervised identity matching method based on embeddings that leverages multiple user features. It constructs fact sets based on user features and learns implicit user representations by mapping facts into an embedding space.

Semi-supervised methods combine a small amount of labeled data with a large amount of unlabeled data to improve matching performance through co-training or multi-view learning. HU Z et al. [4] proposed a semi-supervised framework, SSF, for cross-social-network user matching. MAUIL [5] is a more recent semi-supervised method that employs multi-level attribute joint embedding, using unsupervised techniques to extract user attribute features as a supplement to user network topology features. Li C et al. [6] introduced a noise-aware semi-supervised variational user identity linkage model, NSVUIL, which optimizes the embedding process through multi-granularity modeling at the character, word, article, structure, and label levels, enhancing missing data by leveraging adjacent text attributes.

Supervised methods are currently the mainstream approach, relying on large amounts of labeled data to achieve high-precision matching through trained classifiers or regression models, though data annotation costs are high. Earlier classic models often focused solely on user network topology. For example, the IONE model [7] uses user follow relationships as input and output context vectors, while the

ABNE model [8] employs attention mechanisms. In subsequent research, combining user attribute features with network topology features became more common. Zhang J et al. [9] proposed the MEgo2Vec model, which uses convolutional neural networks to extract character-level and word-level embeddings of user attributes and employs three attention mechanisms to obtain structural embeddings, which are then concatenated for matching. Huang S et al. [10] proposed a bidirectional GCN-based network user identification model that uses the Jaccard similarity coefficient to calculate user intimacy and constructs adjacency matrices to accurately represent relationships in user social networks, though it does not utilize user attribute features. Despite progress, existing methods still face challenges in handling data sparsity and dynamicity. Additionally, platform isolation can lead to discrepancies in the social structures of the same user across platforms. For instance, even if two matched users have a social relationship on one platform, they may not have established a connection on another, resulting in the loss of node features.

To address these issues, this paper proposes a novel supervised cross-social-network user matching method, SSUM, which integrates user attributes and social structures into a unified framework and incorporates a subgraph extraction module to enhance social structure features. Furthermore, the method preprocesses the dataset to complete user social structures, improving matching accuracy. Experiments on real-world social network datasets demonstrate that the proposed method outperforms comparative approaches.

2. SSUM Method

2.1. Problem Definition

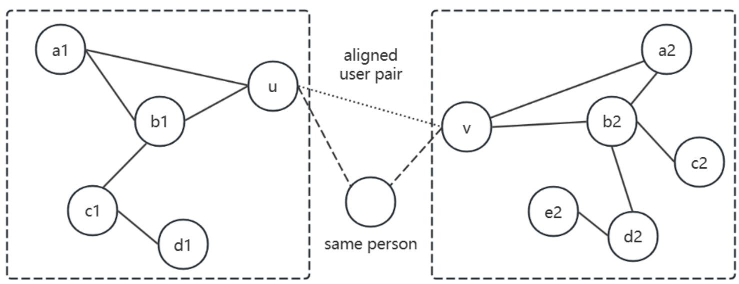

Figure 1. Schematic Diagram of Cross-Social-Network User Matching.

In this method, a social network can be represented as $G = (U, R, X)$, where $U = \{u_1, u_2, u_3 \dots u_n\}$ denotes the set of user nodes across two networks, $R = \{(u_i, u_j) | u_i, u_j \in U\}$ represents the set of relationships between users within the networks, and $X = \{x_{u_1}, x_{u_2}, \dots, x_{u_n}\}$corresponds to the attribute features of the users. Given two social networks $G_1 = (U_1, R_1, X_1)$ and $G_2 = (U_2, R_2, X_2)$, the task of cross-social-network user alignment aims to identify matching user pairs $A = \{(u_i, u_j) | u_i \in U_1, u_j \in U_2\}$ across different social networks.

2.2. Method Overview

This section elaborates on the specifics of the SSUM (Social Structure-based User Matching) method. First, to address the issue of missing user social relationships in the dataset, a preprocessing step is conducted to complete the user social structures. Consider two social networks, G_X and G_Y, where it is known that users U_1^XandU_2^X in network G_X form matching pairs with users U_1^YandU_2^Y in network G_Y. If a relationship$R_1 = \{(U_1^X, U_2^X)|U_1^X, U_2^X \in U^{G_X}\}$ exists in social network G_X, but no corresponding relationship exists between users U_1^YandU_2^Y in social network G_Y, then the missing relationship is constructed and supplemented as $R_2 = \{(U_1^Y, U_2^Y)|U_1^Y, U_2^Y \in U^{G_Y}\}$. This ensures the integrity of the user social structures, thereby facilitating a more effective extraction of user social structural features.

The proposed method employs a multi-dimensional feature integration approach, prioritizing user social structures while supplementing with user attribute information, to achieve cross-social-network user alignment. For the extraction of user attribute features, the method utilizes Word2Vec [11], a widely adopted technique in such tasks, to extract features from short texts such as usernames. For longer user-generated content, a variant of the classic BERT model, namely DistilBERT [12], is employed to ensure a more lightweight architecture and enhance inference speed. Regarding the processing of user social structures, the method introduces a social substructure extraction module. After thoroughly extracting user social substructure features using Graph Neural Networks (GNNs) [13], the extracted substructures, along with the original network, are fed into a Transformer [14] layer to capture global social structures. This process yields the user's social structure features. Finally, the embeddings of user attributes and social structures are concatenated, resulting in a comprehensive and richly attributed user node embedding representation.

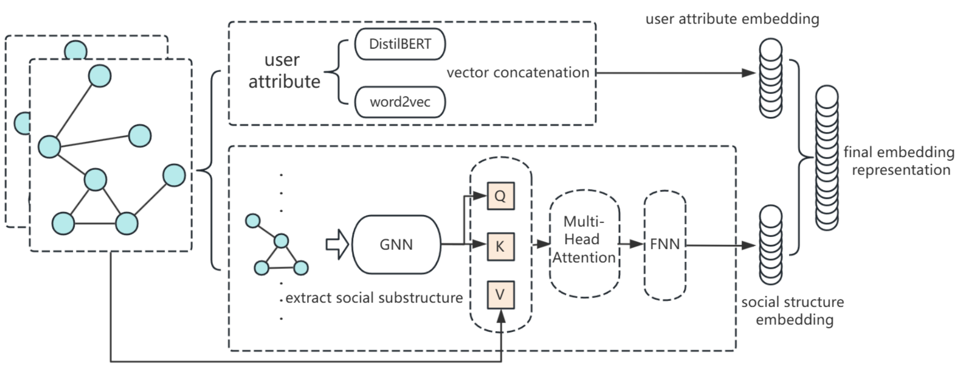

Figure 2. SSUM Method

2.3. Extraction of User Attribute Features

User attributes can generally be divided into two categories. The first category consists of personal profile information, including usernames, gender, addresses, email addresses, and so on. This type of data is characterized by short text length and weak inter-correlation. However, it plays a crucial role in cross-social-network user matching tasks, as users tend to use the same or similar usernames across different social networks. In this method, Word2Vec is employed to extract user profile information, with some adjustments made to suit the task. For short text attributes of users, denoted as $T = \{w_1, w_2, \dots, w_n\}$, where w_i represents the i-th word in the text, word vectors are extracted as follows:

$$v_{w_i} = W[w_i] \tag{1}$$

Subsequently, a weighted average method is applied to aggregate these word vectors, resulting in the embedding representation of the text attribute T:

$$E_{attr} = v_{\text{weighted-avg}} = \frac{\sum_{i=1}^{n} \alpha_i v_{w_i}}{\sum_{i=1}^{n} \alpha_i} \tag{2}$$

where $W[w_i]$ denotes the vector corresponding to the word w_i in the word vector matrix W, and $\alpha_i$represents the weight of each word w_i, which is calculated using the TF-IDF method.

The second category comprises user-generated content. For this, the method directly employs DistilBERT, utilizing its Chinese pre-trained model, distilbert-base-chinese, to extract features. The resulting embedding of the user-generated content is denoted as E_{text}.

2.4. Extraction of User Social Structure Features

The social structure of users plays a critical role in cross-social-network user matching tasks, particularly the relationships with nearby neighbors, while long-distance social relationships contribute less to the matching task. This method focuses on the local social structure of users and introduces a social substructure extraction module to capture the social structure features. The social substructure extractor can be represented as:

$$\varphi(u, G) = \sum_{v \in \mathcal{N}_k(u)} \mathrm{GNN}_G^{(k)}(v) \tag{3}$$

where v represents a node in the network G, and $\varphi(u, G)$ denotes the extraction of the subgraph vector representation centered at u. This method is implemented based on Graph Neural Networks (GNNs), and k represents the k -hop neighbors centered around u. In this method, k is set to 3.

Subsequently, the extracted substructure, along with the original user social structure, is fed into the Transformer layer. The attention mechanism based on the social structure can be expressed as:

$$\mathrm{attn}_{SS}(v) := \sum_{u \in V} \frac{\kappa_{\mathrm{g}}(S_G(v), S_G(u))}{\sum_{w \in V} \kappa_{\mathrm{g}}(S_G(v), S_G(w))} f(x_u) \tag{4}$$

where $\kappa_{\mathrm{g}}(S_G(v), S_G(u))$ represents $\kappa_{\exp}(\varphi(v, G), \varphi(u, G))$, which computes the self-attention between the substructures of two users. Ultimately, we obtain the embedding of the user's social structure as:

$$E_{stru} = x'_v = x_v + \frac{1}{\sqrt{d_v}\,\mathrm{attn}_{SS}(v)} \tag{5}$$

where d_v represents the degree of the node, which to some extent reflects the importance of the user node in the social network.

2.5. Cross-Social-Network User Matching

Finally, we concatenate all the obtained embedding vectors to derive the final user embedding representation:

$$E_{final} = [E_{attr} \parallel E_{text} \parallel E_{stru}] \tag{6}$$

Next, the cross-social-network user matching task is treated as a ranking problem, where users are ranked based on the distance between their node embedding vectors. The model is trained by minimizing the contrastive loss, ensuring that the feature distance between matched user pairs is minimized, while the distance between unmatched user pairs is maximized:

$$L = \frac{1}{N}\sum_{i,j} \left[y_{ij} \cdot d(u_i, v_j)^2 + (1 - y_{ij}) \cdot max(0, \beta - d(u_i, v_j))^2\right] \tag{7}$$

where $d(u_i, v_j)$ represents the Manhattan distance between the user pair (u_i, v_j), β is the margin parameter used to control the minimum distance between unmatched user pairs, and y_{ij} indicates the matching status of the user pair. For y_{ij}, the following holds:

$$y_{ij} = \begin{cases} 1, & (u_i, v_j) \in \rho \\ 0, & (u_i, v_j) \notin \rho \end{cases} \tag{8}$$

where ρ represents the set of matched user pairs.

3. Experiments

In this chapter, we will primarily introduce the datasets used in this study, the design of the experiments, as well as the experimental results and comparisons with other models.

3.1. Data Sets

The two datasets used in the experiments are both freely available, real-world social network datasets. One is a collaboration network dataset of scholars from abroad, while the other is a dataset from domestic social platforms.

- The DBLP17-19 dataset is derived from the DBLP literature database and contains collaboration networks of scholars from two different time periods. In addition to the user attributes and network topology required for cross-social-network user matching tasks, it also includes a wealth of other information, which is ignored during use.
- The Weibo-Douban dataset is a social network dataset crawled from two domestic social platforms, Sina Weibo and Douban. In addition to user attributes and network topology, it also includes user-generated content.

Both datasets, after analysis, exhibit the scale-free property of social networks, meaning that the distribution of node degrees follows a power-law distribution.

Table 1. Dataset Information.

Datasets	#Nodes	#Relations	Min_ Degree	Max_ Degree	Ave_ Degree	#Matched Pairs
DBLP17	9086	56926	2	145	5.7	2832
DBLP19	9325	50155	2	138	5.2	
Weibo	9714	120321	2	652	12.3	1397
Douban	9526	124877	2	608	12.7	

3.2. Experimental Results and Comparative Analysis

3.2.1. Comparative Models

In this paper, we selected the following models for comparison to evaluate the performance of our proposed method:

a) IONE: This model uses the follow and followed-by relationships of users as input and output context vectors, respectively, to achieve network embedding. It then matches users based on their embedding vectors.

b) JARUA [15]: This is a supervised joint embedding method for attributes and relationships. It processes user attributes and relationships separately using different approaches and has demonstrated strong performance.

c) TUMA [16]: This model proposes a graph Transformer-based approach and introduces multiple positional encodings. While it performs well in extracting user network topology features, its overall effectiveness is limited due to its reliance solely on network topology information, without integrating multiple attribute features.

d) ABNE: This is a directed graph-based method for cross-social-network user matching. It employs an attention mechanism for feature extraction but also relies solely on the single attribute of user network topology.

e) MAUIL: This model introduces a novel semi-supervised approach. It uses unsupervised methods to extract three types of text attribute features and user relationships as supplementary feature embeddings, followed by matching. It is also a method that integrates multiple user attributes.

3.2.2. Evaluation Metric

In this paper, we adopt the widely used evaluation metric in cross-social-network user matching tasks, Hit-Precision, which is particularly suitable for ranking-based user matching methods. It measures the proportion of correctly matched users in the candidate list. Specifically, it calculates the ratio of the number of correctly matched users to k in the top k candidate matches. The formula is as follows:

$$Hit - Precision = \frac{1}{|N|}\sum_{x_i \in N} \frac{k-(rank(x_i)-1)}{k} \quad (9)$$

where $rank(x_i)$ returns the position ranking of the correctly matched user among the top k candidate users, and N refers to the total number of successfully matched user pairs. Finally, the average score of the successfully matched user pairs is calculated as the evaluation metric for the model.

3.2.3. Results Analysis

The models compared in this paper can generally be divided into two categories: those using a single attribute and those integrating multiple attributes. The single-attribute methods include IONE, TUMA, and ABNE, while the multi-attribute fusion methods include JARUA and MAUIL. In this experiment, k in the evaluation index was set to three values, namely 1, 3, and 5. By adjusting the value of k, the strictness of the evaluation can be controlled.

Table 2. Comparison of Hit-Precision with Other Methods.

Method	Weibo-Douban			DBLP17-19		
	k=1	k=3	k=5	k=1	k=3	k=5
IONE	0.013	0.023	0.030	0.272	0.311	0.429
ABNE	0.080	0.109	0.118	0.301	0.458	0.487
MAUIL	0.319	0.384	0.412	0.656	0.702	0.736
JARUA	0.416	0.448	0.464	0.721	**0.799**	0.821
TUMA	0.266	0.303	0.359	0.583	0.627	0.678
SSUM	**0.442**	**0.475**	**0.493**	**0.744**	0.786	**0.856**

As shown in the table above, the performance metrics of methods based on multi-attribute fusion are significantly better than those based on a single attribute, demonstrating the effectiveness of combining user attributes with user network topology. Specifically, the proposed method in this paper achieves Hit-Precision scores that are up to 0.176, 0.172, and 0.134 higher than IONE, ABNE, and TUMA, respectively, across the three values of k. This indicates that the proposed method effectively leverages multiple user attributes while emphasizing user social structures, achieving excellent results.

Additionally, the method closest in performance to ours is the JARUA model. However, JARUA relies solely on GAT (Graph Attention Network) to extract user network topology features, which inevitably leads to over-smoothing and over-squeezing issues as the network depth increases, negatively impacting the quality of the extracted features. In contrast, the proposed method combines Transformer with user social substructures to capture long-range dependencies, effectively mitigating the effects of over-smoothing and over-squeezing. Experimental results show that the proposed method outperforms JARUA in most cases across the two datasets for various values of k.

4. Conclusions and Future Work

This paper proposes a social structure-based method for cross-social-network user identity matching, termed SSUM, which effectively identifies the same real-world user across two social networks, alleviating challenges posed by data sparsity and heterogeneity to some extent. The SSUM method begins by preprocessing the dataset to complete the users' social structures. It then introduces a social substructure extraction module to enhance the feature extraction of users' local neighborhood structures, while combining Transformer to capture global network topology information. The integration of these two components yields the user's social structure embedding. Additionally, SSUM employs different feature extraction methods based on text types to obtain user attribute embeddings. Finally, all user attribute embeddings are concatenated to produce the final user embedding, enabling the user matching task. Experimental results demonstrate that this method significantly enhances performance metrics. However, although the method partially addresses the issue of incomplete user social structures, the problem of insufficient labeled user pairs remains. Future work will consider incorporating iterative training methods to expand the number of labeled users in the dataset, thereby further improving the model's performance.

References

[1] Gao, Ming, Lim, Ee Peng, Lo, David, et al. CNL: Collective Network Linkage Across Heterogeneous Social Platforms[C]. In: IEEE International Conference on Data Mining.2015.757-762.https://doi.org/10.1109/ICDM.2015.34

[2] Li, Chaozhuo, Wang, Senzhang, Yu, Philip S, et al. Distribution distance minimization for unsupervised user identity linkage[C]. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management.2018.447-456. https://doi.org/10.1145/3269206.3271675

[3] Tao Zhou,Ee-Peng Lim,Roy Ka-Wei Lee,Feida Zhu,Jiuxin Cao.Retrofitting Embeddings for Unsupervised User Identity Linkage[C].In Proceeding ofthe 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD 2020).Singapore.May 2020.https://doi.org/10.1007/978-3-030-47426-3_30

[4] HU Z, WANG J, CHEN S, et al. A semi-supervised framework with efficient feature extraction and network alignment for user identity linkage[C]//International Conference on Database Systems for Advanced Applications. Cham, Springer, 2021: 675-691. https://doi.org/10.1007/978-3-030-73197-7_46

[5] Baiyang C ,Xiaoliang C .MAUIL: Multilevel attribute embedding for semisupervised user identity linkage[J].Information Sciences,2022,593527-545. https://doi.org/10.1016/j.ins.2022.02.023

[6] Li C , Wang S , Liu Z ,et al.Semi-Supervised Variational User Identity Linkage via Noise-Aware Self-Learning[J]. 2021. https://doi.org/10.48550/arXiv.2112.07373

[7] Liu, L.; Cheung, W.K.; Li, X.; Liao, L. Aligning users across social networks using network embedding. In Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, NY, USA, 9-15 July 2016; pp. 1774-1780. https://dl.acm.org/doi/abs/10.5555/3060832.3060869

[8] Liu, L.; Zhang, Y.M.; Fu, S.; Zhong, F.J.; Hu, J.; Zhang, P. ABNE: An Attention-Based Network Embedding for User Alignment Across Social Networks. IEEE Access 2019, 7, 23595-23605. https://doi.org/10.1109/ACCESS.2019.2900095.

[9] Zhang J , Chen B , Wang X ,et al.MEgo2Vec: Embedding Matched Ego Networks for User Alignment Across Social Networks[J].ACM, 2018.DOI:10.1145/3269206.3271705.

https://doi.org/10.1145/3269206.3271705.

[10] Huang S , Xiang H , Leng C ,et al.Cross-Social-Network User Identification Based on Bidirectional GCN and MNF-UI Models[J].Electronics (2079-9292), 2024, 13(12).DOI:10.3390/electronics13122351. https://doi.org/10.3390/electronics13122351.

[11] Mikolov T , Chen K , Corrado G ,et al.Efficient Estimation of Word Representations in Vector Space[J].Computer Science, 2013.DOI:10.48550/arXiv.1301.3781. https://doi.org/10.48550/arXiv.1301.3781

[12] Sanh V , Debut L , Chaumond J ,et al.DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J]. 2019.DOI:10.48550/arXiv.1910.01108. https://doi.org/10.48550/arXiv.1910.01108

[13] Xu K , Hu W , Leskovec J ,et al.How Powerful are Graph Neural Networks?[J]. 2018.DOI:10.48550/arXiv.1810.00826. https://doi.org/10.48550/arXiv.1810.00826

[14] Vaswani A , Shazeer N , Parmar N ,et al.Attention Is All You Need[J].arXiv, 2017.DOI:10.48550/arXiv.1706.03762. https://doi.org/10.48550/arXiv.2407.15516

[15] Min Y ,Baiyang C ,Xiaoliang C .JARUA: Joint Embedding of Attributes and Relations for User Alignment across Social Networks[J].Applied Sciences,2022,12(24):12709-12709. https://doi.org/10.3390/app122412709.

[16] Lei T ,Ji L ,Wang G , et al.Transformer-Based User Alignment Model across Social Networks[J].Electronics,2023,12(7). https://doi.org/10.3390/electronics12071686.